REMARKS

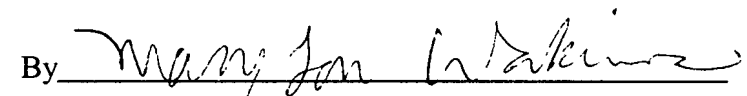
The foregoing amendments to the specification correct grammar, typographical errors, and clarity in terms. For example, the term "non-coding regions" is clarified to include introns and intergenic regions. Support for this is found at least on Specification page 6, lines 9-11 and page 14, lines 3-5 as originally filed. As such, the foregoing amendments to Specification pages 2 and 14 and Claims 4 and 10 (in part) are for consistency. Acceptance is respectfully requested.

Furthermore, the amendments to Claims 4 and 10 are to clarify claim terms and are supported by Specification pages 13-14 and Equation 3, as originally filed. Acceptance is respectfully requested.

No new matter is introduced by this Preliminary Amendment.

Respectfully submitted,

HAMILTON, BROOK, SMITH & REYNOLDS, P.C.

By 

Mary Lou Wakimura
Registration No. 31,804
Telephone: (978) 341-0036
Facsimile: (978) 341-0136

Concord, MA 01742-9133
Dated: 12/21/01

MARKED UP VERSION OF AMENDMENTS

Specification Amendments Under 37 C.F.R. § 1.121(b)(1)(iii)

Replace the paragraph at page 1, lines 4 through 12 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

[A cell has] Most eukaryotic cells have an operational center called the nucleus which contains structures called chromosomes. Chemically, chromosomes are formed of deoxyribonucleic acid (DNA) and associated protein molecules. [Structurally, each] Each chromosome may have tens of thousands of genes. Some genes are referred to as "encoding" (or carrying information for constructing) proteins which are essential in the structuring, functioning and regulating of cells, tissues and organs. Thus, for each organism, the components of the DNA molecules encode [all] much of the information necessary for creating and maintaining life of the organism. See Human Genome Program, U.S. Department of Energy, "Primer on Molecular Genetics", Washington, D.C., 1992.

Replace the paragraph at page 2, lines 9 through 16 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

As previously mentioned, each DNA molecule contains many genes. A gene is a specific sequence of nucleotide bases. These sequences carry the information required for constructing proteins. A protein is a large molecule formed of one or more chains of amino acids in a specific order. Order is determined by base sequence of nucleotides in the gene coding for the protein. Each protein has a [unique function] well-defined functionality. A DNA sequence consists of many biologically distinct regions. For the purpose of this application, Applicants distinguish between intergenic DNA and genes. Many of the genes in mammalian cells are "split genes". A split gene consists of coding and non-coding sequences. The coding sequences in a gene are contained within exonic regions (exons), that appear sequentially separated by long regions referred to as introns. [In a DNA molecule, there are protein-coding sequences (genes) called "exons", and non-coding-

function sequences called "introns" interspersed within many genes. The balance of DNA sequences in the genome are other non-coding regions or intergenic regions.]

Replace the paragraph at page 3, line 3 through 25 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

Gene identification and gene discovery in newly sequenced genomic sequences is one of the most timely computational questions addressed by bioinformatics scientists. Popular gene finding systems include Glimmer, Genmark, Genscan, Genie, GENEWISE, and Grail (See Burge, C. and S. Karlin, "Prediction of complete gene structures in human genomic DNA," *J. Mol. Biol., 268:*78-94, 1997; Salzberg, S. et al., "Microbial gene identification using interpolated Markov models," *Nucl. Acids Res., 26(2)*:544-548, 1998; Xu, Y. et al., "Grail: A multi-agent neural network system for gene identification," *Proc. of the IEEE, 84(10)*:1544-1552, 1996; Kulp, D. et al., "A generalized hidden Markov model for the recognition of human genes in DNA," in *ISMB-96: Proc. Fourth Intl. Conf. Intelligent Systems for Molecular Biology*, pp. 134-141, Menlo Park, CA, 1996, AAAI Press; Borodovsky, M. and J.D. McIninch, "Genemark: Parallel gene recognition for both DNA strands," *Computers & Chemistry, 17(2)*:123-133, 1993; and Salzberg, S. et al. eds., *Computational Methods in Molecular Biology*, Vol. 32 of *New Comprehensive Biochemistry*, Elsevier Science B.V., Amsterdam, 1998). The annotations produced by gene finding systems have been made available to the public. Such projects include the genomes of over thirty microbial organisms, as well as Malaria, Drosophila, C.elegans, mouse, Human [chromosome 22] and others. For instance, Glimmer has been widely used in the analysis of many microbial genomes and has reported over 98% accuracy in prediction accuracy (See Fraser, C.M. et al., "Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi," *Nature 390(6660)*:580-586, December 1997). Genie (D. Kulp et al. above) has been deployed in the analysis of the Drosophila genome, and Genscan (C. Burge and S. Karlin above) was used for analysis of human chromosome 22.

Replace the paragraph at page 4, lines 13 through 23 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

On a very high level, genes in human DNA and many other organisms have a relatively simple structure. All eukaryotic genes, including human genes, are thought to share a similar layout. This layout adheres to the following "grammar" or pattern: start codon, exon, $(\text{intron-exon})^n$, stop codon. The start codon is a specific 3-base sequence (e.g. ATG) which signals the beginning of the gene. Exons [are] contain the actual genetic material that code for proteins as mentioned above. Introns are the spacer segments of DNA whose function is not clearly understood. [And finally] Finally, stop codons (e.g., TAA) [which] signal the end of the gene. The notation $(\text{intron-exon})_n$ simply means that there are n alternating intron-exon segments. Genes identification procedures has to take into account other important issues such as polyA tail, promoters, pseudo-genes, alternative splicing and other features.

Replace the paragraphs at page 5, lines 3 through 7 and lines 8 through 16 with the below paragraphs marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

General theory for the combination of experts has drawn significant interest in the machine learning community. Theory and practice of combining experts have been studied in literature. [Some examples are the] The choice of a particular way of combining expert predictions depends on the properties of individual experts and the demands posed by the problem at hand.

Most techniques for combining gene predictions proposed in the past have been rather trivial or have relied on *ad hoc* combinations of experts. In one prior [work] project, Murakami and Takagi (Murakami, K. and T. Takagi, "Gene recognition by combination of several gene-finding programs," *Bioinformatics, 14(8)*:665-675, 1998) proposed a system for gene recognition that combines several gene-finding programs. They implemented an AND and OR combination, HIGHEST-method (best individual expert), RULE-method (decisions using sets of expert rules), and an *ad hoc* BOUNDARY-method. The best of these methods achieved an improvement in general accuracy of 3%-5% over the individual gene finders.

Replace the paragraph at page 6, lines 8 through 11 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

In accordance with another aspect of the present invention, the [The] preliminarily predicted gene locations and/or predicted genes include exon (or coding regions) predictions. Alternatively, the gene locations for predicted genes are indicated by exons and introns (i.e., coding and non-coding regions) of the subject genome sequence.

Replace the paragraph at page 7, lines 3 through 9 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

By way of background, ligated exons [are the sequence regions that are translated into proteins] form a sequence that is translated into a protein by a simple but still computationally mysterious mechanism of splicing that takes place after the DNA sequence has been transcribed into RNA. The process starts by spliceome proteins that recognize the splice signals, followed by a step where the introns are cut out (spliced out), and ending in a phase where the consecutive exons are "glued" together into a single sequence that is translated into a protein. Intuitively speaking this process is performed on an RNA "image" of the genomic sequence.

Replace the paragraphs at page 8, lines 1 through 5 and lines 6 through 8 with the below paragraphs marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

The present invention is a system for the combination of individual experts which is learned from data. Unlike the prior art, such a system exploits learned dependencies between experts and forms a prediction maximally consistent with known gene data. Statistically, predictions of the invention system [will then] have the potential to [generalize to genes undiscovered by any of the individual experts] refine the boundaries and verify the predictions made by experts.

An attractive [way of combining experts which] methodology that exploits [their] the joint [statistical behavior and can thus satisfy requirements of] statistics of expert systems and avoids the shortcomings of the prior art[,] is based on Bayesian networks.

Replace the paragraph at page 9, lines 3 through 15 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

Referring back to Fig. 1A, the Bayesian network combiner 17 is trained using known DNA sequences with known genes (exons/introns) as the training data 35. The training data is applied to the computer apparatus. That is, a corresponding character string representing the known DNA sequence is input to the invention software 21. Invention software 21 applies this training data character string to the expert systems 15. The expert systems 15 each determine/predict preliminary exons/introns 19 in the training data 35. The preliminary exon/intron predictions 19 from the expert systems 15 are fed into the Bayesian network combiner 17. In turn, the Bayesian network combiner 17 combines the preliminary exon/intron predictions 19 in a manner consistent with the known genes (exons/introns locations and pattern). That is, the Bayesian network combiner 17 is [adjusted] trained to make the combination of preliminary exon/intron predictions produce the known exons on output. In this way, the Bayesian network combiner 17 is said to be trained on the training data 35.

Replace the paragraph at page 10, lines 7 through 18 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

Bayesian network probabilistic models provide a flexible and powerful framework for statistical inference as well as learning of model parameters from data. The goal of inference is to find a distribution of [a] one or more random variables in the network conditioned on evidence (known values) of other variables. Bayesian networks encompass efficient inference algorithms, such as Jensen's junction tree (Jensen, F.V., *An Introduction to Bayesian Networks*, Spring-Verlag, 1995) or Pearl's message passing (Pearl, J., *Probabilistic reasoning in intelligent systems*, Morgan Kaufmann, San Mateo, CA 1998). Inside a learning loop, such algorithms may be used to efficiently estimate optimal values of a model's parameters from data (for instance, see Jordan, M.I. ed., *Learning in Graphical Models*, Kluwer Academic Publishers, 1998). Furthermore, techniques exist that can optimally determine the topology of a Bayesian network together with its parameters directly from data.

Replace the paragraph at page 11, lines 9 through 18 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

Gene combiner parameters, probability tables $P(E_i|Y)$ and $P(Y)$, are learned from a training dataset of nucleotide sequences by statistically calculating $P(E_i|Y)$ and $P(Y)$ of all individual predictors E_i and labeled for ground truth Y. For instance, a maximum likelihood (ML) estimate of these parameters for a training set of N nucleotides is

$$P(E_i = e|Y = y) = \frac{\#E_i = e, Y = y}{N}$$

where e denotes the prediction of an expert system i, e ϵ {intron,exon}, and y is the combined prediction, y ϵ {intron,exon}. # E_i=e, Y=y denotes the number of cases in the training dataset where the prediction of expert system i is e and the [ground truth] <u>combined prediction</u> is y. Alternative estimates of these parameters may be obtained using MAP (maximum a posteriori) estimation.

Replace the paragraph at page 12, lines 12 through 16 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

It can be easily shown that the performance of the full Bayes model 31 is at least as good as that of the best individual expert system 25, 27, 29. Furthermore, the [often] <u>previously</u> used AND, OR and majority models are special cases of the full Bayes combiner 31. Nevertheless, this model 31 still assumes that the annotation of a particular nucleotide is independent of the annotation of any other nucleotide in the sequence.

Replace the paragraph at page 12, lines 22 through 24 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

Index t in Figure 4 refers to the sample at location t in the subject nucleotide sequence. For instance, the sequence Y^{t-1},Y^{t},Y^{t+1} gives the combined annotation for the subject nucleotide sequence <u>n positions t-1, t, and t+1 respectively</u>.

Replace the paragraph at page 14, lines 6 through 13 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

For that purpose, Applicants assumed that each individual expert system provides the following binary decision. An expert system produces a single labeling for every nucleotide in a sequence: E if the nucleotide is a part of an exon and I if it belongs to an intron or an intergenic region. Using the notation of Applicants' models, $E_i \in \{E, I\}$ for an expert i. Similarly, a combined decision Y is either E or I. Parameters of each of the four above-discussed models of Bayesian network combiners 28, 31, 40, 51 were learned using a standard maximum likelihood estimation in the Bayesian network framework. All prediction results were then obtained using a five-fold cross-validation.

Replace the paragraph at page 15, lines 7 through 12 with the below paragraph marked up by way of bracketing and underlining to show the changes relative to the previous version of the paragraph.

An exon is said to be exactly predicted 47 only if both its ending and beginning points coincides with that of a true exon. An exon is said to be missed 57 if there is no overlap with any of the predicted exons. ME gives the percentage of missed exons 57 whereas WE gives the percentage of wrongly or overpredicted exons 49. To compute these two numbers (ME and WE), Applicants look for any overlap between a true and a predicted exon. Wrong exon (WE) prediction implies the prediction has no overlap with a true exon.

Claim Amendments Under 37 C.F.R. § 1.121(c)(1)(ii)

4. (Amended) Computer apparatus as claimed in Claim 3 wherein the Bayesian network combines the predicted gene locations according to

$$Y^* = \max_{Y_t} P(Y_t|E_1,\ldots,E_n, Y^*_{t-1})$$

$E_i \in \{E, I\}$

where t is location in the subject genomic sequence and $E_1,...,E_n$ are the respective [predicted gene locations of] predictions (E for exon or I for intron or intergenic region) made by individual units 1 through n, n being the number of units in the plurality.

10. (Amended) A method as claimed in Claim 9 wherein the step of combining using a Bayesian network combines according to

$$Y^* = \max_{Y_t} P(Y_t | E_1,...,E_n, Y^*_{t-1})$$

$$E_i \in \{E, I\}$$

where t is location in the subject genomic sequence and $E_1,...,E_n$ are the respective [predicted gene locations of] predictions (E for exon or I for intron or intergenic region) made by individual expert systems, n being the number of expert systems in the plurality.